

CAI  
BS 1  
-1989  
R24

**MAINFRAME SAS ENHANCEMENTS IN THE SUPPORT  
OF EXPLORATORY DATA ANALYSIS**

by

Richard Johnson and Jane F. Gentleman

No. 24

Statistics Canada  
Analytical Studies Branch

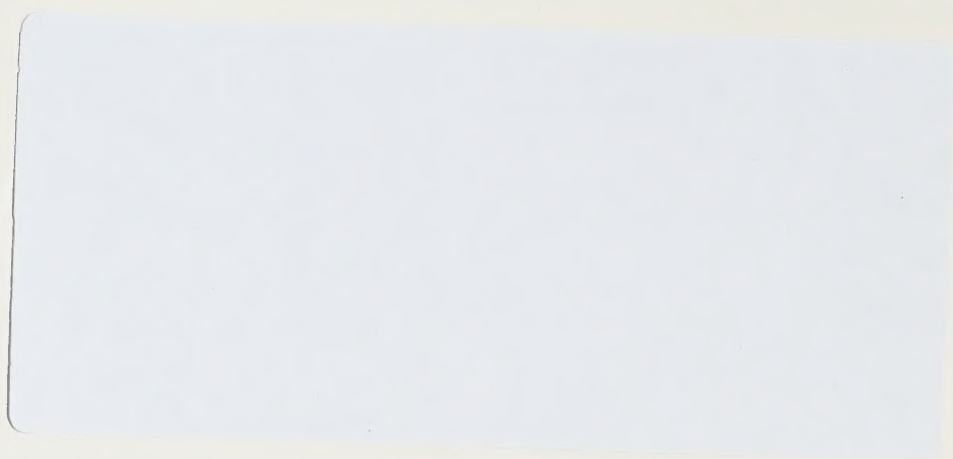
# Research Paper Series



Statistics  
Canada

Statistique  
Canada

Canada



CAI  
BSI  
-1989  
R24

**MAINFRAME SAS ENHANCEMENTS IN THE SUPPORT  
OF EXPLORATORY DATA ANALYSIS**

by

Richard Johnson and Jane F. Gentleman

No. 24

Social and Economic Studies Division  
Analytical Studies Branch  
Statistics Canada  
1989

The analysis presented in this paper is the responsibility of the authors and does not necessarily represent the views or policies of Statistics Canada.

Aussi disponible en français



# Mainframe SAS Enhancements in Support of Exploratory Data Analysis

by Richard Johnson and Jane F. Gentleman


## ABSTRACT

This document is a manual describing computer software developed for exploratory data analysis at Statistics Canada. The new software comprises a collection of SAS functions and macros, with heavy emphasis on graphics. Together with some functions already available in the SAS system, these routines perform the following types of operations: evaluation of probability density functions (PDF's), cumulative distribution functions (CDF's), and inverse CDF's for nine distributions; calculation of sample quantiles; generation of random numbers; graphing of histograms with optional PDF superimposition; calculation of empirical CDF's with optional graphing and optional CDF superimposition; and graphing of Q-Q and P-P plots comparing a sample to any of the nine distributions. Detailed instructions are provided for using the software to construct Q-Q plots comparing two samples.

Key Words: Exploratory data analysis  
Graphics

Received: May 5, 1989

Accepted: October 10, 1989



Digitized by the Internet Archive  
in 2023 with funding from  
University of Toronto

<https://archive.org/details/31761103748422>

## Foreword

This document is a manual describing computer software for exploratory data analysis, the results of a joint effort between members of the Informatics Branch and the Analytical Studies Branch at Statistics Canada. The manual is being distributed by the Informatics Branch to the computer user community within Statistics Canada, and it also appears here as a paper in the Analytical Studies Branch Research Paper Series. It is being re-published under the latter auspices because the Analytical Studies Branch Research Paper Series is intended to represent the broad array of activities being carried out within the Branch, including the production of research papers eventually destined for refereed journal publication, as well as other types of activities which support research and analysis.

The software described herein was originally developed for use by students in a data analysis class offered by Statistics Canada as part of its on-going effort to increase data analytic capability within the agency. The software is now available for general use by Statistics Canada personnel. The Informatics Branch will release future updates of the manual as the software is further enhanced.

Future enhancements for this software will include incorporation of weights for sample data and additional distributions.

Following the manual are examples of graphs produced by the macros.

It is not the intention of the Informatics Branch to distribute the software outside Statistics Canada. However, those who wish may request a photocopy of the new source code from the second author.

Steven Earwaker coordinated and managed the data analysis class for which this software was produced. Louise Bergeron provided expert programming assistance.

Richard Johnson, Informatics Branch

Jane F. Gentleman, Analytical Studies Branch



## PREFACE

# Mainframe SAS Enhancements in Support of Exploratory Data Analysis

SAS Support Staff

Informatics Branch  
Statistics Canada

July 20, 1989



## PREFACE

This document describes SAS functions and macros developed specifically in support of the pilot presentation of Statistics Canada course 04181 entitled "The Art of Data Analysis". These facilities have been implemented on the Statistics Canada mainframe computer system. It is assumed that the reader is generally familiar with the mainframe operating environment and SAS Version 5 as configured for that environment.

Eighteen functions were written to supplement those already available in the SAS system. When viewed together with nine of the original SAS functions, they constitute a collection of routines to evaluate three types of statistical probability functions: probability density functions (PDF's), cumulative distribution functions (CDF's), and inverse cumulative distribution functions. These theoretical functions are useful tools in data analysis. In addition, macros have been developed to perform more complex operations involving both the theoretical distributions and data. In the following list, some uses for the theoretical functions are given, and operations for which macros have been developed are identified.

- Plots of PDF's can be compared to appropriately constructed histograms. A SAS macro has been developed to produce superimposed plots of this nature.
- CDF's are useful for calculating significance levels (P-values), calculating certain goodness-of-fit test statistics, calculating coordinates for P-P plots, and comparing theoretical to empirical cumulative distribution functions (ECDF's). A SAS macro has been developed to compute and plot ECDF's and optionally superimpose CDF's. Another macro has been developed to produce P-P plots.
- Inverse CDF's can be used to calculate theoretical quantiles, calculate coordinates for Q-Q plots, and generate random numbers. A SAS macro has been developed to produce Q-Q plots.
- A SAS macro for calculating sample quantiles has been developed. This is useful for analyzing the distribution of a sample of data and for calculating coordinates for Q-Q plots.
- SAS macros have been developed, based on available SAS random number functions, to facilitate the generation of random numbers for five distribution types.

Sections 1 and 2 of this document provide detailed descriptions of how to use the SAS functions and macros. Section 3 describes the minimal operational considerations associated with the production of graphics. An Appendix contains formulas for the probability density functions.

The pilot version of "The Art of Data Analysis" was designed and presented by Dr. Jane Gentleman of Social and Economic Studies Division, Analytical Studies Branch. Coordination

and logistics were managed by Stephen Earwaker of Social Survey Methods Division, Methodology Branch.

This document and the special SAS facilities described herein were created by the SAS support staff of Informatics Branch, Statistics Canada. The author of this document wishes to acknowledge: the University of Waterloo, upon whose Fortran algorithms were based some of the SAS functions written in support of the course; Dr. Gentleman and Stephen Earwaker for valuable advice provided throughout the software development process; and Dr. Gentleman for assistance in the writing and editing of this document.

At this early stage in the anticipated life of the functions and macros, it is possible that some errors may be present in the routines or in this document. Please notify the designated SAS Software Representative in Informatics Branch of any suspected errors. (The SAS help library member SITEINFO contains current information regarding SAS support contacts.) Specifications for the use of these facilities may change with continued development and refinement.

# CONTENTS

<b>Preface</b> . . . . .	<b>ii</b>
 <b>Section 1: Probability Functions</b> . . . . .	 <b>1</b>
PDF's, CDF's and Inverse CDF's for Nine Distributions . . . . .	1
Probability Density Functions Written at Statistics Canada . . . . .	3
PDFBETA . . . . .	3
PDFCHI . . . . .	3
PDFEXP . . . . .	3
PDFF . . . . .	4
PDFGAM . . . . .	4
PDFNORM . . . . .	4
PDFT . . . . .	5
PDFUNI . . . . .	5
PDFWEI . . . . .	5
Cumulative Distribution Functions Written at Statistics Canada . . . . .	6
PROBEXP . . . . .	6
PROBUNI . . . . .	6
PROBWEI . . . . .	7
Inverse Cumulative Distribution Functions Written at Statistics Canada . . . . .	7
CHIINV . . . . .	7
EXPINV . . . . .	7
FINV . . . . .	8
TINV . . . . .	8
UNIINV . . . . .	8
WEIINV . . . . .	9
 <b>Section 2: Macros</b> . . . . .	 <b>10</b>
Histograms with Optional PDF Superimposition . . . . .	11
%HIST Macro . . . . .	12
ECDF's with Optional Plotting and Optional CDF Superimposition . . . . .	14
%ECDF Macro . . . . .	14
Sample Quantiles . . . . .	16
%QUANT Macro . . . . .	16
Probability (Q-Q and P-P) Plots . . . . .	17
Quantile-Quantile (Q-Q) Plots Comparing a Sample to a Distribution . . . . .	17
%QQ Macro . . . . .	18
Probability-Probability (P-P) Plots Comparing a Sample to a Distribution . . . . .	19
%PP Macro . . . . .	20

Quantile-Quantile (Q-Q) Plots Comparing Two Samples . . . . .	21
Random Variate Generators . . . . .	22
%GENCHI Macro . . . . .	22
%GENEXP Macro . . . . .	23
%GENNORM Macro . . . . .	23
%GENT Macro . . . . .	24
%GENUNI Macro . . . . .	25
 <b>Section 3: User Interface . . . . .</b>	 <b>26</b>
Batch Mode . . . . .	26
Interactive Mode . . . . .	26
 <b>Appendix A: Formulas for Probability Density Functions . . . . .</b>	 <b>28</b>
Beta Distribution . . . . .	28
Chi-square Distribution . . . . .	28
Exponential Distribution . . . . .	28
F Distribution . . . . .	29
Gamma Distribution . . . . .	29
Normal Distribution . . . . .	29
t Distribution . . . . .	29
Uniform Distribution . . . . .	30
Weibull Distribution . . . . .	30

## TABLES

1. A Comprehensive List of Probability Functions . . . . .	2
--	---

## Section 1

### PROBABILITY FUNCTIONS

The SAS system provides a variety of probability functions as documented in Chapter 6 of "SAS User's Guide: Basics, Version 5 Edition". Numerous additional functions have been written, using VS Fortran, to complement those provided by SAS and complete a set viewed as desirable to meet the objectives of the course "The Art of Data Analysis".

#### ***1.1 PDF's, CDF's and Inverse CDF's for Nine Distributions***

The comprehensive set of relevant functions consists of three functions for each of nine distribution types. The functions are: the probability density function (PDF); the cumulative distribution function (CDF); and the inverse of the cumulative distribution function (inverse CDF). Table 1 lists the nine distributions covered, gives the name of each function, and indicates whether the function is an original SAS function or was written at Statistics Canada.

This document describes the functions written at Statistics Canada. "SAS User's Guide: Basics" is the appropriate reference for the original SAS functions. Like SAS's own functions, the locally written functions are routines that return a value computed from one or more arguments. They are used in the context of a DATA step and, typically, are executed with each iteration of the DATA step, that is, as the DATA step processes each observation in a SAS data set. Arguments to the functions documented below are positional and mandatory. They are subjected to range validation. If invalid or missing arguments are detected, messages will be written to the SAS log and results will be set to missing.

Table 1: A Comprehensive List of Probability Functions

Distribution	Parameters	Functions		
		PDF	CDF	Inverse CDF
Normal	mu,sigsq	PDFNORM	PROBNORM <sup>1,2</sup>	PROBIT <sup>1,3</sup>
Uniform	a,b	PDFUNI	PROBUNI	UNIINV
Exponential	theta	PDFEXP	PROBEXP	EXPINV
t	df	PDFT	PROBT <sup>1</sup>	TINV
Chi square	df	PDFCHI	PROBCHI <sup>1</sup>	CHIINV
F	df1,df2	PDFF	PROBF <sup>1</sup>	FINV
Weibull	lo,sc,sh	PDFWEI	PROBWEI	WEIINV
Gamma	lo,sc,sh	PDFGAM	PROBGAM <sup>1,4</sup>	GAMINV <sup>1,5</sup>
Beta	a,b	PDFBETA	PROBBETA <sup>1</sup>	BETAINV <sup>1</sup>

Notes:

1. Original SAS function. (All others written at Statistics Canada.)
2. PROBNORM( $x$ ) is the CDF of a  $N(0,1)$  random variable at argument  $x$ . Use PROBNORM( $(x-\mu)/\sigma$ ) for the CDF of a  $N(\mu,\sigma^2)$  random variable at argument  $x$  (where  $\mu = \text{mu}$  and  $\sigma^2 = \text{sigsq}$ ).
3. PROBIT( $P$ ) is the inverse CDF of a  $N(0,1)$  random variable at argument  $P$ . Use  $(\sigma)\text{PROBIT}(P)+\mu$  for the inverse CDF of a  $N(\mu,\sigma^2)$  random variable at argument  $P$  (where  $\mu = \text{mu}$  and  $\sigma^2 = \text{sigsq}$ ).
4. PROBGAM( $x,sh$ ) is the CDF of a  $\text{Gamma}(0,1,sh)$  random variable at argument  $x$ . Use PROBGAM( $(x-lo)/sc,sh$ ) for the CDF of a  $\text{Gamma}(lo,sc,sh)$  random variable at argument  $x$ .
5. GAMINV( $P,sh$ ) is the inverse CDF of a  $\text{Gamma}(0,1,sh)$  random variable at argument  $P$ . Use  $(sc)\text{GAMINV}(P,sh)+lo$  for the inverse CDF of a  $\text{Gamma}(lo,sc,sh)$  random variable at argument  $P$ .

## 1.2 Probability Density Functions Written at Statistics Canada

### 1.2.1 PDFBETA

The PDFBETA function returns the probability density at argument  $x$  of a beta distribution with parameters  $a$  and  $b$ .

General form:

**PDFBETA( $x, a, b$ )**

where

- x**                    The value at which the probability density is to be evaluated.  $0 < x < 1$ .
- a**                    First shape parameter.  $a > 0$ .
- b**                    Second shape parameter.  $b > 0$ .

### 1.2.2 PDFCHI

The PDFCHI function returns the probability density at argument  $x$  of a Chi-square distribution with  $df$  degrees of freedom.

General form:

**PDFCHI( $x, df$ )**

where

- x**                    The value at which the probability density is to be evaluated.  $x > 0$ .
- df**                   Number of degrees of freedom.  $df \geq .5$ . Argument  $df$  need not be an integer.

### 1.2.3 PDFEXP

The PDFEXP function returns the probability density at argument  $x$  of an exponential distribution with mean  $theta$ .

General form:

**PDFEXP( $x, theta$ )**

where

- x**                    The value at which the probability density is to be evaluated.  $x \geq 0$ .
- theta**                Mean  $theta$ .  $Theta > 0$ .

### 1.2.4 PDFF

The PDFF function returns the probability density at argument  $x$  of an F distribution with  $df1$  and  $df2$  degrees of freedom.

General form:

**PDFF( $x, df1, df2$ )**

where

- |            |  |
|------------|--|
| <b>x</b>   | The value at which the probability density is to be evaluated. $x > 0$ .           |
| <b>df1</b> | Numerator degrees of freedom. $df1 > 0$ . Argument $df1$ need not be an integer.   |
| <b>df2</b> | Denominator degrees of freedom. $df2 > 0$ . Argument $df2$ need not be an integer. |

### 1.2.5 PDFGAM

The PDFGAM function returns the probability density at argument  $x$  of a gamma distribution with the given *location*, *scale* and *shape* parameters.

General form:

**PDFGAM( $x, lo, sc, sh$ )**

where

- |           |   |
|-----------|---|
| <b>x</b>  | The value at which the probability density is to be evaluated. $x > lo$ . |
| <b>lo</b> | <i>Location</i> parameter. $-\infty < lo < \infty$ .                      |
| <b>sc</b> | <i>Scale</i> parameter. $sc > 0$ .  |
| <b>sh</b> | <i>Shape</i> parameter. $sh > 0$ .  |

### 1.2.6 PDFNORM

The PDFNORM function returns the probability density at argument  $x$  of a Normal distribution with mean  $\mu$  and variance  $\text{sigsq}$ .

General form:

**PDFNORM( $x, \mu, \text{sigsq}$ )**

where

- |          |   |
|----------|---|
| <b>x</b> | The value at which the probability density is to be evaluated. $-\infty < x < \infty$ . |
|----------|---|

**mu**                Mean ( $\mu$ ).  $-\infty < \mu < \infty$ .

**sigsq**            Variance ( $\sigma^2$ ).  $\sigma > 0$ .

### 1.2.7    PDFT

The PDFT function returns the probability density at argument  $x$  of a  $t$  distribution with  $df$  degrees of freedom.

General form:

**PDFT(x,df)**

where

**x**                The value at which the probability density is to be evaluated.  $-\infty < x < \infty$ .

**df**               Number of degrees of freedom.  $df > 0$ . Argument  $df$  need not be an integer.

### 1.2.8    PDFUNI

The PDFUNI function returns the probability density at argument  $x$  of an uniform distribution on the interval  $[a,b]$ .

General form:

**PDFUNI(x,a,b)**

where

**x**                The value at which the probability density is to be evaluated.  $a \leq x \leq b$ .

**a**                Lower limit of the interval.  $a < b$ .

**b**                Upper limit of the interval.  $a < b$ .

### 1.2.9    PDFWEI

The PDFWEI function returns the probability density at argument  $x$  of a Weibull distribution with the given *location*, *scale* and *shape* parameters.

General form:

**PDFWEI(x,lo,sc,sh)**

where

<b>x</b>	The value at which the probability density is to be evaluated. $x > lo$ .
<b>lo</b>	<i>Location</i> parameter. $-\infty < lo < \infty$ .
<b>sc</b>	<i>Scale</i> parameter. $sc > 0$ .
<b>sh</b>	<i>Shape</i> parameter. $sh > 0$ .

### 1.3 Cumulative Distribution Functions Written at Statistics Canada

#### 1.3.1 PROBEXP

The PROBEXP function returns the probability that a random variable having the exponential distribution with mean *theta* is less than or equal to the input argument *x*.

General form:

**PROBEXP(x,theta)**

where

<b>x</b>	The value at which the function is to be evaluated. $x \geq 0$ .
<b>theta</b>	Mean <i>theta</i> . $Theta > 0$ .

#### 1.3.2 PROBUNI

The PROBUNI function returns the probability that a random variable having the uniform distribution on the interval  $(a,b)$  is less than or equal to the input argument *x*.

General form:

**PROBUNI(x,a,b)**

where

<b>x</b>	The value at which the function is to be evaluated. $a \leq x \leq b$ .
<b>a</b>	Lower limit of the interval. $a < b$ .
<b>b</b>	Upper limit of the interval. $a < b$ .

### 1.3.3 PROBWEI

The PROBWEI function returns the probability that a random variable having the Weibull distribution with the given *location*, *scale* and *shape* parameters is less than or equal to the input argument *x*.

General form:

**PROBWEI(*x*,*lo*,*sc*,*sh*)**

where

- x**                    The value at which the function is to be evaluated.  $x > lo$ .
- lo**                    *Location* parameter.  $-\infty < lo < \infty$ .
- sc**                    *Scale* parameter.  $sc > 0$ .
- sh**                    *Shape* parameter.  $sh > 0$ .

## 1.4 *Inverse Cumulative Distribution Functions Written at Statistics Canada*

### 1.4.1 CHIINV

The CHIINV function returns the Chi-square value *x*, such that a random variable, distributed as Chi-square with *df* degrees of freedom, is less than or equal to *x* with probability *p*.

General form:

**CHIINV(*p*,*df*)**

where

- p**                    Probability in range [0,1].
- df**                    Number of degrees of freedom.  $df \geq .5$ . Argument *df* need not be an integer.

### 1.4.2 EXPINV

The EXPINV function returns the exponential value *x*, such that a random variable, distributed as exponential with mean *theta*, is less than or equal to *x* with probability *p*.

General form:

**EXPINV(*p*,*theta*)**

where

<b>p</b>	Probability in range [0,1].
<b>theta</b>	Mean <i>theta</i> . <i>Theta</i> > 0.

### 1.4.3 FINV

The FINV function returns the F value  $x$ , such that a random variable, distributed as F with  $df1$  and  $df2$  degrees of freedom, is less than or equal to  $x$  with probability  $p$ .

General form:

**FINV(p,df1,df2)**

where

<b>p</b>	Probability in range [0,1].
<b>df1</b>	Numerator degrees of freedom. $df1 > 0$ . Argument $df1$ need not be an integer.
<b>df2</b>	Denominator degrees of freedom. $df2 > 0$ . Argument $df2$ need not be an integer.

### 1.4.4 TINV

The TINV function returns the t value  $x$ , such that a random variable, distributed as t with  $df$  degrees of freedom, is less than or equal to  $x$  with probability  $p$ .

General form:

**TINV(p,df)**

where

<b>p</b>	Probability in range [0,1].
<b>df</b>	Number of degrees of freedom. $df > 0$ . Argument $df$ need not be an integer.

### 1.4.5 UNIINV

The UNIINV function returns the uniform value  $x$ , such that a random variable, distributed as uniform on the interval  $(a,b)$ , is less than or equal to  $x$  with probability  $p$ .

General form:

**UNIINV(p,a,b)**

where

<b>p</b>	Probability in range [0,1].
<b>a</b>	Lower limit of the interval. $a < b$ .
<b>b</b>	Upper limit of the interval. $a < b$ .

#### 1.4.6 WEIINV

The WEIINV function returns the Weibull value  $x$ , such that a random variable, distributed as Weibull with the given *location*, *scale* and *shape* parameters, is less than or equal to  $x$  with probability  $p$ .

General form:

**WEIINV(p,lo,sc,sh)**

where

<b>p</b>	Probability in range [0,1].
<b>lo</b>	<i>Location</i> parameter. $-\infty < lo < \infty$ .
<b>sc</b>	<i>Scale</i> parameter. $sc > 0$ .
<b>sh</b>	<i>Shape</i> parameter. $sh > 0$ .



## Section 2

### MACROS

This section describes SAS macros written to perform operations on entire SAS data sets. The macros consist of macro statements, and data step programming statements and/or complete DATA and PROC steps. All of the macros were written for name-style macro calls; that is, the form of the invocation is `%macroname(parameters)` as described in Chapter 19 of "SAS User's Guide: Basics, Version 5 Edition".

Several of the macros described in this section produce plots. All plots are produced in landscape orientation on the IBM 3800-3 laser page printer. This device has been chosen because it is available to all mainframe SAS users, and also because plots sent directly to the 3800-3 are produced in the preferred landscape orientation. (By comparison, landscape orientation of graphs on a cut-sheet 3820 laser printer requires creation of a graphics catalog, followed by template replay to effect 90-degree rotation.)

Other graphics devices can be used with the macros described in this section only by modifying the GOPTIONS statement in the macro source code. For occasional needs, this can more easily be done on-line under Display Manager where the macro can be brought into the editor screen with an INCLUDE INCL(macroname) command (where *macroname* is the name of the macro but without the leading % character), modified and SUBMITTED. Once a modified macro has been submitted from the editor screen, that version will take precedence over the original version in the default autocall library whenever that macro is invoked subsequently in the SAS session. In batch mode, the user must create a modified copy of the macro in a user autocall library and then specify the root of that library to the INCL parameter of the JCL procedure. The batch approach can also be used on-line and is appropriate for regular use of an alternative graphics device such as a pen plotter or graphics display terminal. Please note that some of the plotting macros described below have subordinate macros which do the actual plotting and therefore contain the GOPTIONS statement. In the macro descriptions which follow (in the narrative portion preceding the detailed specifications), the name of the macro which contains the GOPTIONS statement is given.

By default, SAS/GRAPH will automatically scale the axes of a plot by taking into consideration the ranges of values of the variables being plotted, the character cell alignment of tick-mark values on the axes, and space required for user-specified titles and footnotes. If several data samples having differing value ranges are plotted in this way, the axes of the plots will have different limits<sup>1</sup> and therefore may be given different lengths even though titling may be consistent and the same graphics device is being used. Such differences in the plots may hinder comparison of the samples. Fortunately, SAS/GRAPH provides the means to force axis

---

<sup>1</sup> In this discussion of axes, the term *length* will refer to the physical length (measurable in inches or centimeters), and the term *limits* will refer to the numerical length reflected by the tick-mark values.

consistency so that plots of differing samples can be compared.

Thus, each of the plotting macros can operate in either of two modes regarding axis length and limits, depending on the use of available parameters to the macros. Axis length can be determined by SAS (parameter `AXES=SAS` or no `AXES` parameter specified) or can be fixed at a predetermined percentage<sup>2</sup> of the dimensions of the total plotting surface (`AXES=FIXED`). Fixed mode imposes restrictions on the use of titles and footnotes, the restrictions varying with the choice of graphics device because of differing character cell dimensions. On the 3800-3 used by the macros, titles and footnotes of default height can be used in 2-and-1 combinations: either 2 titles and one footnote, or 2 footnotes and one title. With no footnotes, a maximum of four titles of default height can be used. Heights other than the defaults may be used on a trial-and-error basis.

The axis limits can be allowed to default to those of the relevant variable, or can be specified via `XAXIS` and `YAXIS` parameters. Values specified for the `XAXIS` and `YAXIS` parameters can correspond to any of the forms values can take for the `ORDER` option of the `SAS/GRAPH AXIS` statement. Values can be specified as a list (e.g., `XAXIS=1 3 5 7 9`), as a range (e.g., `XAXIS=1 9` or `XAXIS=1 TO 9`), as a range with an increment (e.g., `XAXIS=1 TO 9 BY 2`), or as a combination of any of these forms. In the context of these macros, specification of the parameter value as a range with an increment will generally be the most appropriate. In this case, SAS will attempt to annotate the axis tick-marks with the specified incremental values. Regardless of the method used to specify axis values, the values will always be evenly spaced along the relevant axis.

## 2.1 *Histograms with Optional PDF Superimposition*

A macro has been written to produce a histogram wherein the widths of the vertical bars can vary in accordance with user-specified boundaries. The macro will also create an output SAS data set having one observation for each histogram interval and containing variables for lower interval boundary, upper interval boundary, frequency and bar height. Optionally, the macro will superimpose a user-specified theoretical probability density function (PDF) between user-specified lower and upper limits. The graph will be produced on the 3800-3 laser printer. To use an alternative graphics device, it is necessary to modify the `GOPTIONS` statement contained in the `%HIST` macro. (See the general information at the beginning of this section.)

In order to scale the histogram properly so that a PDF can be superimposed, the bar heights are calculated as follows:

$$\text{bar height} = \frac{\text{number of data values in the interval}}{(\text{interval width})(\text{total number of data values})}$$

Thus, like a PDF, the total area of the histogram is one.

Parameters must be provided to identify the input SAS data set and to specify interval boundaries for the histogram. Any data value which is equal to a boundary between two bars is counted as being in the higher interval. A data value equal to the lowest/highest boundary

---

<sup>2</sup> The percentage is fixed in the macro code and cannot be modified by the user without macro modification.

is counted as being in the leftmost/rightmost interval. If any data values fall outside the outer histogram boundaries specified by the user, the macro will generate additional lower and/or upper histogram intervals to accommodate the data.

If the appropriate parameter is provided to select a PDF, then a graph of that function will be superimposed on the histogram. In this case, the user has the option of specifying lower and upper plotting limits for the PDF via another parameter, or letting the minimum and maximum sample values be used by default.

Macro parameters are in the form *keyword=value* and can be specified in any order.

### 2.1.1 %HIST Macro

General form:

**%HIST(parameter list)**

Parameters:

- IN=** the name of the input sample SAS data set for which a histogram is to be produced. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate.
- INVAR=** the name of the variable, in the input SAS data set identified by the IN= parameter, for which a histogram is to be produced. If this parameter is not specified, the variable name X will be assumed.
- BNDS=** the interval boundaries for the histogram. Specify the lower bound for each bar, proceeding from left to right, and terminate the list with the upper bound of the rightmost bar. (There are no gaps between bars.) Separate the boundary values by at least one blank. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate.
- AXES=** the method used to determine the lengths of the axes for the histogram. AXES=SAS will allow SAS to determine the lengths of the axes. AXES=FIXED will fix the axes at a predetermined percentage of the total plotting surface dimensions. If this parameter is not specified, AXES=SAS will be assumed. (See also the discussion of axis length at the beginning of this section.)
- XAXIS=** the X-axis limits and intermediate tick-mark values. This parameter is optional. It is less likely to be used in this macro than in other plotting macros. This is because %HIST simulates a histogram through a plotting technique. As a result, SAS/GRAPH provides evenly spaced X-axis tick-marks that do not necessarily correspond to user specified interval boundaries. (PROC GPLOT doesn't know that it is drawing a histogram.) If XAXIS values corresponding to interval boundaries are specified, the boundaries will be evenly spaced along the X-axis, thereby defeating the purpose of the macro if the intervals are not of equal width. (See the discussion of axis values at the beginning of this section for details on the syntax of this parameter.)

- YAXIS=** the Y-axis limits and intermediate tick-mark values. This parameter is optional. If used, its value can correspond to any of the forms values can take for the **ORDER** option of the **SAS/GRAPH AXIS** statement. (See the discussion of axis values at the beginning of this section for details on the syntax of this parameter.)
- FUNC=** the PDF specification which will result in a theoretical curve being superimposed on the histogram. This parameter is optional. If it is used it must be specified in the form **FUNC=***function-name(arguments)*, where *function-name* is the name of a PDF described in Section 1.2 of this document, and *arguments* is the argument list required by the chosen function. The first element in the function's argument list, which is the name of the variable at which the function is to be evaluated, must always be coded as **X**. (This is not the same **X** as the default **INVAR** variable.)
- LIMITS=** the minimum and maximum values of **X** (argument to the PDF) at which the PDF will be evaluated. (The macro will generate 200 values of **X**, evenly distributed between the limits. The PDF will be evaluated and plotted as 200 points connected by straight lines.) The limits must be given in the order **LIMITS=***lower upper* with at least one blank separator between the limits. This parameter will be used only if **FUNC=** has been specified. If **FUNC=** is specified and **LIMITS=** is not specified, then the minimum and maximum sample values will be used by default. If the lower limit specified for the PDF is greater than the minimum sample value, then the latter will be used instead as the lower limit. Similarly, if the upper limit specified for the PDF is less than the maximum sample value, then the latter will be used instead as the upper limit.
- OUT=** the name of the output SAS data set which will contain information about the intervals of the histogram. The data set will contain one observation for each interval. The observations will contain the following variables:
- |               |  |
|---------------|--|
| <b>LOWER</b>  | the lower boundary for an interval of the histogram.   |
| <b>UPPER</b>  | the upper boundary for an interval of the histogram.   |
| <b>FREQ</b>   | the frequency with which the value range described by the boundaries occurs in the input sample. |
| <b>HEIGHT</b> | the height of the histogram bar which depicts the current interval.                              |

This parameter is required in order to obtain an output data set. If it is not specified, no output data set will be produced.

The following example will produce a histogram of the variable **SAMPL** in SAS data set **WORK.NORM** using interval boundaries as specified. The histogram will be plotted as 9 bars beginning at -2.5 and ending at 3.5. It will be overlayed with a plot of the function **PDFNORM** for 200 values evenly distributed between -3.6 and 3.6. Arguments to **PDFNORM** are: the mandatory variable name **X**, mean 0, and variance 1. An output SAS

data set named HISTINFO will be created. It will contain nine observations, one for each interval.

```
%HIST(IN=NORM, INVAR=SAMPL,  
      BNDS=-2.5 -2 -1.5 -1 -.25 .25 1.5 2 2.75 3.5,  
      FUNC=PDFNORM(X,0,1),LIMITS=-3.6 3.6,  
      OUT=HISTINFO)
```

## 2.2 ECDF's with Optional Plotting and Optional CDF Superimposition

A macro has been written to calculate the empirical cumulative distribution function (ECDF) evaluated at all the ordered observations of a designated variable. The ECDF is defined as follows. If  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  are the sample values sorted in non-descending order, then the value of the ECDF at argument  $x_{(i)}$  is  $(i-.5)/n$ .

The results of the calculation are placed in a new variable which is written to an output data set along with all input variables. An input data set must be specified in the parameter list. If an output data set is not specified, the input data set will be reused. It should be noted that, as a result of a sort step in the macro, the output data set will be produced in ascending sequence of the input variable. If the original sequence of the input data set must be retained, it is necessary to specify an output data set name in the parameter list.

Optionally, a plot of the ECDF will be produced on the 3800-3 laser printer. In order to use an alternative graphics device, it is necessary to modify the GOPTIONS statement contained in the %ECDFPLT macro called by %ECDF. (See the general information at the beginning of this section.) If an ECDF plot is requested, a user-specified theoretical cumulative distribution function (CDF) can optionally be superimposed on the plot of the ECDF by specifying the appropriate parameter to select a CDF. In this case, the user has the option of specifying lower and upper plotting limits for the CDF via another parameter, or letting the minimum and maximum sample values be used by default.

Macro parameters are in the form *keyword=value* and can be specified in any order.

### 2.2.1 %ECDF Macro

General form:

```
%ECDF(parameter list)
```

Parameters:

**IN=** the name of the input SAS data set. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate.

**INVAR=** the name of the input variable for which the ECDF is to be calculated. If this parameter is not specified, the variable name X will be assumed.

**NEWVAR=** the name of the new variable which will contain the ECDF values. If this parameter is not specified, the name `F_HAT` will be used.

**OUT=** the name of the output SAS data set to be produced. If this parameter is not specified, a warning message will be written to the SAS log and the input data set will be reused.

**PLOT=** indicates whether or not an ECDF plot is to be produced on the 3800-3 laser printer. The default value is NO. Specify `PLOT= YES` to obtain a plot.

**AXES=** the method used to determine the lengths of the axes for the optional plot. `AXES=SAS` will allow SAS to determine the lengths of the axes. `AXES=FIXED` will fix the axes at a predetermined percentage of the total plotting surface dimensions. If this parameter is not specified, `AXES=SAS` will be assumed. If `PLOT=NO` is in effect, this parameter will be ignored. (See also the discussion of axis length at the beginning of this section.)

**XAXIS=** the X-axis limits and intermediate tick-mark values. This parameter is optional. If used, its value can correspond to any of the forms values can take for the `ORDER` option of the `SAS/GRAPH AXIS` statement. If `PLOT=NO` is in effect, this parameter will be ignored. (See the discussion of axis values at the beginning of this section for details on the syntax of this parameter.)

**YAXIS=** the Y-axis limits and intermediate tick-mark values. (See `XAXIS=.`)

**FUNC=** the CDF specification which will result in a theoretical curve being superimposed on the ECDF plot produced in response to the `PLOT= YES` parameter. (`PLOT= YES` is, therefore, a prerequisite to the use of this parameter.) This parameter must be given in the form `FUNC=function-name(arguments)`, where *function-name* is the name of a CDF described either in "SAS User's Guide: Basics, Version 5 Edition" or in Section 1.3 of this document, and *arguments* is the argument list required by the chosen function. The first element in the function's argument list, which is the name of the variable at which the function is to be evaluated, must always be coded as X. (This is not the same X as the default `INVAR` variable.)

**LIMITS=** the minimum and maximum values of X (argument to the CDF) at which the CDF will be evaluated. (The macro will generate 200 values of X, evenly distributed between the limits. The CDF will be evaluated and plotted as 200 points connected by straight lines.) The limits must be given in the order `LIMITS=lower upper` with at least one blank separator between the limits. This parameter will be used only if `FUNC=` has been specified. If `FUNC=` is specified and `LIMITS=` is not specified, then the minimum and maximum sample values will be used by default. If the lower limit specified for the CDF is greater than the minimum sample value, then the latter will be used instead as the lower limit. Similarly, if the upper limit specified for the CDF is less than the maximum sample value, then the latter will be used instead as the upper limit.

## 2.3 Sample Quantiles

A macro has been written to compute sample quantiles. The sample P-quantile  $Q(P)$  is defined as follows. Let  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  be the sample values sorted in non-descending order. Then  $Q(P) = x_{(nP+.5)}$ , where linear interpolation is used if  $1 < nP+.5 < n$  and  $nP+.5$  is not an integer. If  $P < .5/n$ , then  $Q(P) = x_{(1)}$ . If  $P > (n-.5)/n$ , then  $Q(P) = x_{(n)}$ . The quantity  $nP+.5$  is the "index" of  $Q(P)$ .

Two SAS data sets are required as input to this macro: a sample data set containing the variable from which quantiles are to be computed, and a data set containing one or more probabilities for which quantiles are to be computed. The SAS data set containing the probabilities will consist of one observation for each probability. Parameters must be provided for both input data sets. The macro produces a sorted, temporary copy of the input sample data set; the sequence of the original is not altered. An output data set is created containing one observation for each observation in the data set of probabilities. Output observations will contain the input probability variable and computed quantile and quantile-index variables. Note that if no parameter is provided to name the output data set, the input data set of probabilities will be reused. In this case, all original variables will be retained and the quantile and quantile-index variables will be added. Macro parameters are in the form *keyword=value* and can be specified in any order.

### 2.3.1 %QUANT Macro

General form:

**%QUANT(parameter list)**

Parameters:

- IN=** the name of the input sample SAS data set from which quantiles are to be computed. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate.
- INVAR=** the name of the sample variable, in the input SAS data set identified by the IN= parameter, from which quantiles are to be computed. If this parameter is not specified, the variable name X will be assumed.
- INP=** the name of the input SAS data set containing one or more probabilities for which quantiles are to be computed. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate.
- INPVAR=** the name of the variable, in the input SAS data set identified by the INP= parameter, containing the probabilities for which quantiles are to be computed. These probabilities may be any values between zero and one and need not be sorted. If this parameter is not specified, a message will be written to the SAS log and the macro will terminate.
- OUT=** the name of the output SAS data set which will contain the computed quantiles. The data set will contain one observation for each observation in the input

probabilities data set. If this parameter is not specified, the input probabilities data set, specified by the INP= parameter, will be reused. In this case, all variables contained in the INP= data set will be retained.

**Q=** the name of the output variable which will contain the computed quantiles. If this parameter is not specified, the variable name Q will be used.

**QI=** the name of the output variable which will contain the computed quantile indices. If this parameter is not specified, the variable name QI will be used.

## 2.4 Probability (Q-Q and P-P) Plots

Macros have been written to construct two types of probability plots for comparing a sample of data to a theoretical distribution: (1) Quantile-Quantile (Q-Q) plots to compare sample quantiles to theoretical quantiles, and (2) Probability-Probability (P-P) plots to compare sample cumulative proportions to theoretical cumulative probabilities. In addition, instructions are given below for constructing a Q-Q plot comparing two samples of data to each other.

### 2.4.1 Quantile-Quantile (Q-Q) Plots Comparing a Sample to a Distribution

Let  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  be the sample values sorted in non-descending order. Let  $P_1, P_2, \dots, P_k$  be a set of  $k$  probabilities selected by the user; these may be any values between zero and one and need not be sorted. Let  $Q_1, Q_2, \dots, Q_k$  be the corresponding sample quantiles. (The  $Q_i$ 's are calculated from the entire sample of  $n$  values.)

A Q-Q plot comparing the sample to a theoretical distribution is a scatter plot of the  $k$  points  $(F^{-1}(P_i), Q_i)$  (for  $i = 1, \dots, k$ ), where  $F^{-1}$  is the inverse CDF for the desired theoretical distribution (see Section 1).

The values  $k = n$  and  $P_i = (i-.5)/n$  are commonly used, in which case  $Q_i = x_{(i)}$ .

The macro for Q-Q plotting will operate in one of two modes, depending on whether or not the user provides input probabilities. In either case, a SAS data set containing an input sample is required.

If input probabilities are not provided, the ECDF will be calculated, using the %ECDF macro, at each observation of the input sample. The result is used as the probability argument to an inverse CDF function which the user must specify as a parameter to the macro. Values returned by the inverse CDF function are used as X-coordinates for the Q-Q plot. The input sample values are used as Y-coordinates.

If input probabilities are provided, the %QUANT macro is used to obtain P-quantiles from the the input sample and these are used as the Y-coordinates. The user-specified inverse CDF function operates on the same P-quantiles to produce the X-coordinates.

An optional output SAS data set can be produced containing the coordinates for the Q-Q plot. The plot will be produced on the 3800-3 laser printer. In order to use an alternative

graphics device, it is necessary to modify the GOPTIONS statement contained in the %PORQ macro called by %QQ. (See the general information at the beginning of this section.) Macro parameters are in the form *keyword=value* and can be specified in any order.

#### 2.4.1.1 %QQ Macro

General Form:

**%QQ(parameter list)**

Parameters:

- IN=** the name of the input sample SAS data set. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate. The input sample SAS data set need not be sorted by the user.
- INVAR=** the name of the variable which contains the input sample. If this parameter is not specified, the variable name X will be assumed.
- INP=** the name of the SAS data set containing the optional input probabilities. The macro will operate in one of two modes depending on whether or not this parameter has been specified. (See text preceding the macro parameter specifications for details.) If this parameter is used, the INPVAR= parameter must also be given.
- INPVAR=** the name of the variable containing the input probabilities. This parameter is required if the INP= parameter has been specified. If INP= has been specified and INPVAR= has not, a message will be written to the SAS log and the macro will terminate.
- FUNC=** the inverse CDF specification for the desired theoretical distribution. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate. This function will be evaluated either at each of the default probability values derived from the ECDF calculation, or at each value of the INPVAR variable, depending on the operational mode of the macro, in order to produce theoretical quantiles to be used as X-axis coordinates for the Q-Q plot. This parameter must be given in the form *FUNC=function-name(arguments)*, where *function-name* is the name of an inverse CDF described either in "SAS User's Guide: Basics, Version 5 Edition" or in Section 1.4 of this document, and *arguments* is the argument list required by the chosen function. The first element in the function's argument list, which is the name of the variable at which the function is to be evaluated, must always be coded as X.
- AXES=** the method used to determine the lengths of the axes for the plot. **AXES=SAS** will allow SAS to determine the lengths of the axes. **AXES=FIXED** will fix the axes at a predetermined percentage of the total plotting surface dimensions. If this parameter is not specified, **AXES=SAS** will be assumed. (See also the discussion of axis length at the beginning of this section.)

- XAXIS=** the X-axis limits and intermediate tick-mark values. This parameter is optional. If used, its value can correspond to any of the forms values can take for the **ORDER** option of the **SAS/GRAPH AXIS** statement. (See the discussion of axis values at the beginning of this section for details on the syntax of this parameter.)
- YAXIS=** the Y-axis limits and intermediate tick-mark values. (See **XAXIS=**.)
- OUT=** the name of the optional output SAS data set which will contain the X and Y coordinates for the Q-Q plot. The output data set will be produced only if this parameter has been specified.
- QX=** the name to be given to the output variable which will contain the theoretical quantiles used as X-axis coordinates for the Q-Q plot. If this parameter is specified without a value for the **OUT=** parameter, then it will be ignored. If **OUT=** has been specified and this parameter has not, then the default variable name **QX** will be used.
- QY=** the name to be given to the output variable which will contain the sample quantiles used as Y-axis coordinates for the Q-Q plot. If this parameter is specified without a value for the **OUT=** parameter, then it will be ignored. If **OUT=** has been specified and this parameter has not, then the default variable name **QY** will be used.

## 2.4.2 Probability-Probability (P-P) Plots Comparing a Sample to a Distribution

Let  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  be the sample values sorted in non-descending order. Let  $Q_1, Q_2, \dots, Q_k$  be a set of  $k$  quantile values selected by the user; these need not be sorted. Let  $P_1, P_2, \dots, P_k$  be the corresponding values of the ECDF at the arguments  $Q_i$ . (The ECDF is calculated from the entire sample of  $n$  values.)

A P-P plot comparing the sample to a theoretical distribution is a scatter plot of the  $k$  points  $(F(Q_i), P_i)$  (for  $i = 1, \dots, k$ ), where  $F$  is the CDF for the desired theoretical distribution (see Section 1).

The values  $k = n$  and  $Q_i = x_{(i)}$  are commonly used, in which case  $P_i = (i - .5)/n$ .

The macro for P-P plotting will operate in one of two modes, depending on whether or not the user provides input quantiles. In either case, a SAS data set containing an input sample is required.

If input quantiles are not provided, the ECDF will be calculated, using the **%ECDF** macro, at each observation of the input sample. The results will be used as the Y coordinates for the P-P plot. A CDF function which the user must specify as a parameter to the macro will also be evaluated at each observation of the input sample. Values returned by the CDF will be used as the X coordinates.

If input quantiles are provided, the macro will use linear interpolation to calculate the corresponding probability values of the ECDF at each quantile argument and the results will be used as Y coordinates to the P-P plot. The user-specified CDF will be evaluated at each quantile argument to get the X coordinates.

An optional output SAS data set can be produced containing the coordinates for the P-P plot. The plot will be produced on the 3800-3 laser printer. In order to use an alternative graphics device, it is necessary to modify the GOPTIONS statement contained in the %PORQ macro called by %PP. (See the general information at the beginning of this section.) Macro parameters are in the form *keyword=value* and can be specified in any order.

### 2.4.2.1 %PP Macro

General Form:

**%PP(parameter list)**

Parameters:

- IN=** the name of the input sample SAS data set. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate. The input sample SAS data set need not be sorted by the user.
- INVAR=** the name of the variable which contains the input sample. If this parameter is not specified, the variable name X will be assumed.
- INQ=** the name of the SAS data set containing the optional input quantile values. The macro will operate in one of two modes depending on whether or not this parameter has been specified. (See text preceding the macro parameter specifications for details.) If this parameter is used, the INQVAR= parameter must also be given.
- INQVAR=** the name of the variable containing the input quantile values. This parameter is required if the INQ= parameter has been specified. If INQ= has been specified and INQVAR= has not, a message will be written to the SAS log and the macro will terminate.
- FUNC=** the CDF specification for the desired theoretical distribution. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate. This function will be evaluated either at each value of the input sample, or at each value of the INQVAR variable, depending on the operational mode of the macro, in order to produce theoretical probabilities to be used as X-axis coordinates for the P-P plot. This parameter must be given in the form *FUNC=function-name(arguments)*, where *function-name* is the name of a CDF described either in "SAS User's Guide: Basics, Version 5 Edition" or in Section 1.3 of this document, and *arguments* is the argument list required by the chosen function. The first element in the function's argument list, which is the name of the variable at which the function is to be evaluated, must always be coded as X.

- AXES=** the method used to determine the lengths of the axes for the plot. **AXES=SAS** will allow SAS to determine the lengths of the axes. **AXES=FIXED** will fix the axes at a predetermined percentage of the total plotting surface dimensions. If this parameter is not specified, **AXES=SAS** will be assumed. (See also the discussion of axis length at the beginning of this section.)
- XAXIS=** the X-axis limits and intermediate tick-mark values. This parameter is optional. If used, its value can correspond to any of the forms values can take for the **ORDER** option of the **SAS/GRAPH AXIS** statement. (See the discussion of axis values at the beginning of this section for details on the syntax of this parameter.)
- YAXIS=** the Y-axis limits and intermediate tick-mark values. (See **XAXIS=.**)
- OUT=** the name of the optional output SAS data set which will contain the X and Y coordinates for the P-P plot. The output data set will be produced only if this parameter has been specified.
- PX=** the name to be given to the output variable which will contain the theoretical probabilities used as X-axis coordinates for the P-P plot. If this parameter is specified without a value for the **OUT=** parameter, then it will be ignored. If **OUT=** has been specified and this parameter has not, then the default variable name **PX** will be used.
- PY=** the name to be given to the output variable which will contain the sample probabilities used as Y-axis coordinates for the P-P plot. If this parameter is specified without a value for the **OUT=** parameter, then it will be ignored. If **OUT=** has been specified and this parameter has not, then the default variable name **PY** will be used.

### 2.4.3 Quantile-Quantile (Q-Q) Plots Comparing Two Samples

Let  $x_1, x_2, \dots, x_n$  and  $y_1, y_2, \dots, y_m$  be two samples of data. (These need not be the same size, and they need not be sorted.) Let  $P_1, P_2, \dots, P_k$  be a single set of  $k$  probabilities selected by the user. Let  $Q_1^x, Q_2^x, \dots, Q_k^x$  be the corresponding quantiles for the  $x$ 's, obtained by using the %QUANT macro, and let  $Q_1^y, Q_2^y, \dots, Q_k^y$  be the corresponding quantiles for the  $y$ 's, obtained by using the %QUANT macro a second time.

Then a Q-Q plot comparing the distribution of the  $x$ 's to that of the  $y$ 's is a scatter plot of the  $k$  points  $(Q_i^x, Q_i^y)$  (for  $i = 1, \dots, k$ ).

In the special case where  $n = m = k$  and  $P_i = (i - .5)/n$ , then the Q-Q plot is simply a scatter plot of the ordered  $y$ 's versus the ordered  $x$ 's.

## 2.5 Random Variate Generators

Several macros have been written to facilitate the creation of SAS data sets containing a random variate. Each macro will generate a different distribution: chi-square, exponential, normal, t, or uniform. Keyword parameters enable the user to specify the desired number of observations, the name of the random variate, the initial seed for the random generator function used, and parameters of the distribution. Parameters are specified in the form *key-word*=*value*, and can be specified in any order.

Each macro is designed to be invoked in the context of a data step for which the user supplies a DATA statement identifying the SAS data set to which the macro will output observations. For example,

```
DATA NORM01;  
  %GENNORM(N=200)  
RUN;
```

will produce SAS data set WORK.NORM01 containing 200 observations of normal random variate X with mean 0 and variance 1 as determined by parameter defaults.

Each of these random variate generator macros uses one of the SAS random number functions (two, in the case of %GENT). In the macro descriptions which follow, pertinent SAS random number functions are identified. The SAS random number functions are described in Chapter 6 of "SAS User's Guide: Basics, Version 5 Edition". Techniques used to generate an observation are indicated therein.

The seed parameter associated with each of these macros becomes the seed to the relevant SAS random number function(s). Your attention is directed to page 236 of "SAS User's Guide: Basics" for a detailed discussion of the initialization of a random number stream.

**Please be advised that, if data generated by these macros are to be reproducible, an initial seed having a value greater than zero must be used and that initial seed value should be recorded.**

### 2.5.1 %GENCHI Macro

%GENCHI generates observations of a Chi-square variate. This macro uses the RANGAM function as follows:

```
ALPHA = degrees-of-freedom / 2;  
variate = 2 * RANGAM(seed,ALPHA);
```

where *degrees-of-freedom*, *variate* and *seed* are macro parameters DF, VAR and S, respectively.

General form:

```
%GENCHI(parameter list)
```

Parameters:

**DF=** degrees of freedom. The default value is 1. Any value specified must be an integer.

**N=** the number of observations to be generated. The default value is 50.

**S=** the initial seed for the random number function. The default value is 0 which causes a CPU clock observation to be used as the initial seed. A reproducible series of values can be obtained by using a seed > 0. (See "SAS User's Guide: Basics, Version 5 Edition", page 236.)

**VAR=** the name of the random variate for which values will be generated. The default value is X.

### 2.5.2 %GENEXP Macro

%GENEXP generates observations of an exponential variate. This macro uses the RANEXP function as follows:

```
variate = RANEXP(seed) * THETA;
```

where *variate*, *seed* and *THETA* are macro parameters VAR, S and THETA, respectively.

General form:

```
%GENEXP(parameter list)
```

Parameters:

**THETA=** the mean. The default value is 1.

**N=** the number of observations to be generated. The default value is 50.

**S=** the initial seed for the random number function. The default value is 0 which causes a CPU clock observation to be used as the initial seed. A reproducible series of values can be obtained by using a seed > 0. (See "SAS User's Guide: Basics, Version 5 Edition", page 236.)

**VAR=** the name of the random variate for which values will be generated. The default value is X.

### 2.5.3 %GENNORM Macro

%GENNORM generates observations of a Normal random variate. This macro uses the RANNOR function as follows:

```
variate = mu + SQRT(sigsq) * RANNOR(seed);
```

where *variate*, *mu*, *sigsq* and *seed* are macro parameters VAR, MU, SIGSQ and S, respectively.

General form:

**%GENNORM(parameter list)**

Parameters:

- MU=** the mean. The default value is 0.
- N=** the number of observations to be generated. The default value is 50.
- S=** the initial seed for the random number function. The default value is 0 which causes a CPU clock observation to be used as the initial seed. A reproducible series of values can be obtained by using a seed > 0. (See "SAS User's Guide: Basics, Version 5 Edition", page 236.)
- SIGSQ=** the variance of the distribution. The default value is 1.
- VAR=** the name of the random variate for which values will be generated. The default value is X.

## 2.5.4 %GENT Macro

%GENT generates observations of a t variate. This macro uses the RANNOR and RANGAM functions as follows:

```
ALPHA = degrees-of-freedom / 2;  
R1 = RANNOR(seed);                /* Normal(0,1) */  
R2 = 2 * RANGAM(seed,ALPHA);      /* Chi-square */  
variate = R1 / SQRT(R2 / degrees-of-freedom); /* t */
```

where *degrees-of-freedom*, *seed* and *variate* are macro parameters DF, S and VAR, respectively. (Note that each function call returns a new value for *seed*, thereby ensuring the independence of the Normal and chi-square random variates.)

General form:

**%GENT(parameter list)**

Parameters:

- DF=** degrees of freedom. The default value is 1. Any value specified must be an integer.
- N=** the number of observations to be generated. The default value is 50.
- S=** the initial seed for the random number function. The default value is 0 which causes a CPU clock observation to be used as the initial seed. A reproducible series of values can be obtained by using a seed > 0. (See "SAS User's Guide: Basics, Version 5 Edition", page 236.)

**VAR=**            the name of the random variate for which values will be generated. The default value is X.

### 2.5.5 %GENUNI Macro

%GENUNI generates observations of a uniform random variate on the interval (0,1). This macro uses the RANUNI function as follows:

```
variate = RANUNI(seed);
```

where *variate* and *seed* are macro parameters VAR and S, respectively.

General form:

```
%GENUNI(parameter list)
```

Parameters:

**N=**            the number of observations to be generated. The default value is 50.

**S=**            the initial seed for the random number function. The default value is 0 which causes a CPU clock observation to be used as the initial seed. A reproducible series of values can be obtained by using a seed  $\neq 0$ . (See "SAS User's Guide: Basics, Version 5 Edition", page 236.)

**VAR=**            the name of the random variate for which values will be generated. The default value is X.

## Section 3

### USER INTERFACE

The libraries containing the macros and functions described by this document are made available automatically to users of the RGS-supported (STC2.SAS) catalogued procedures and clists. (Originally, it was necessary for a user to establish access to the libraries by means of parameters to the procedures and clists, but this requirement has been eliminated prior to distribution of this document.)

Some minor operational requirements which remain pertinent are described below.

#### 3.1 *Batch Mode*

When you intend to use a macro or macro option which will generate a SAS/GRAPH plot, you must invoke the SASG3800 catalogued procedure. This procedure allocates the files necessary for graphics output to IBM laser printers via interface with GDDM<sup>3</sup> software, and allocates a default virtual storage region adequate for most graphics applications which use that interface. Symbolic parameters DEST and GCOPIES can be used to route graphs to a remote 38xx printer (where applicable), and to generate multiple copies of graphs, respectively. (GCOPIES should be used with discretion because graphics images are retransmitted to the printer for each copy.)

#### 3.2 *Interactive Mode*

The default TSO logon region is adequate for most SAS sessions that do not use SAS/GRAPH. However, if graphics are to be produced, the logon region should be set to 3000K as in the following example.

```
TSO userid A(account) S(3000)
```

When you intend to use a macro or macro option which will generate a SAS/GRAPH plot, you must specify the GRAPH38 parameter when you invoke the SAS clist. This is done subsequent to issuing the START SAS command and allocating required personal data sets. The following example illustrates the command sequence.

---

<sup>3</sup> IBM's Graphical Data Display Manager

**START SAS**

...

**ALLOC F(filename) DA('dsname')**

...

**SAS GRAPH38**

Parameters DEST and GCOPIES can be used with the SAS clist to route graphs to a remote 38xx printer (where applicable), and to generate multiple copies of graphs, respectively. (Detailed TSO help information is available for the SAS clist upon return from the START SAS command. GCOPIES should be used with discretion because graphics images are retransmitted to the printer for each copy.)

## Appendix A

### FORMULAS FOR PROBABILITY DENSITY FUNCTIONS

#### A.1 *Beta Distribution*

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

$$0 < x < 1$$

$$a > 0$$

$$b > 0$$

#### A.2 *Chi-square Distribution*

$$f(x) = \frac{1}{2^{\frac{df}{2}} \Gamma\left(\frac{df}{2}\right)} e^{-\frac{x}{2}} x^{\frac{df}{2}-1}$$

$$x > 0$$

$$df \geq .5$$

#### A.3 *Exponential Distribution*

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$$

$$x \geq 0$$

$$\theta > 0 \quad (\theta \equiv \text{theta})$$

#### A.4 *F Distribution*

$$f(x) = \frac{\Gamma\left(\frac{a+b}{2}\right)}{\Gamma\left(\frac{a}{2}\right)\Gamma\left(\frac{b}{2}\right)} \left(\frac{b}{a}\right)^{\frac{b}{2}} x^{\frac{a}{2}-1} \left(x + \frac{b}{a}\right)^{-(a+b)/2}$$

$$x > 0$$

$$a > 0 \quad (a \equiv df1)$$

$$b > 0 \quad (b \equiv df2)$$

#### A.5 *Gamma Distribution*

$$f(x) = \frac{\left(\frac{x-\mu}{\sigma}\right)^{\alpha-1} e^{-\left(\frac{x-\mu}{\sigma}\right)}}{\sigma\Gamma(\alpha)}$$

$$x > \mu \quad (\mu \equiv lo)$$

$$-\infty < \mu < \infty$$

$$\sigma > 0 \quad (\sigma \equiv sc)$$

$$\alpha > 0 \quad (\alpha \equiv sh)$$

#### A.6 *Normal Distribution*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$-\infty < x < \infty$$

$$-\infty < \mu < \infty \quad (\mu \equiv mu)$$

$$\sigma > 0 \quad (\sigma^2 \equiv sigsq)$$

#### A.7 *t Distribution*

$$f(x) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)\sqrt{\pi}\sqrt{df}} \left(1 + \frac{x^2}{df}\right)^{-(df+1)/2}$$

$$-\infty < x < \infty$$

$$df > 0$$

## A.8 Uniform Distribution

$$f(x) = \frac{1}{b-a}$$

$$a \leq x \leq b$$

$$a < b$$

## A.9 Weibull Distribution

$$f(x) = \left(\frac{\alpha}{\sigma}\right) \left(\frac{x-\mu}{\sigma}\right)^{\alpha-1} e^{-\left(\frac{x-\mu}{\sigma}\right)^{\alpha}}$$

$$x > \mu \quad (\mu \equiv lo)$$

$$-\infty < \mu < \infty$$

$$\sigma > 0 \quad (\sigma \equiv sc)$$

$$\alpha > 0 \quad (\alpha \equiv sh)$$



## EXAMPLES OF GRAPHS \*

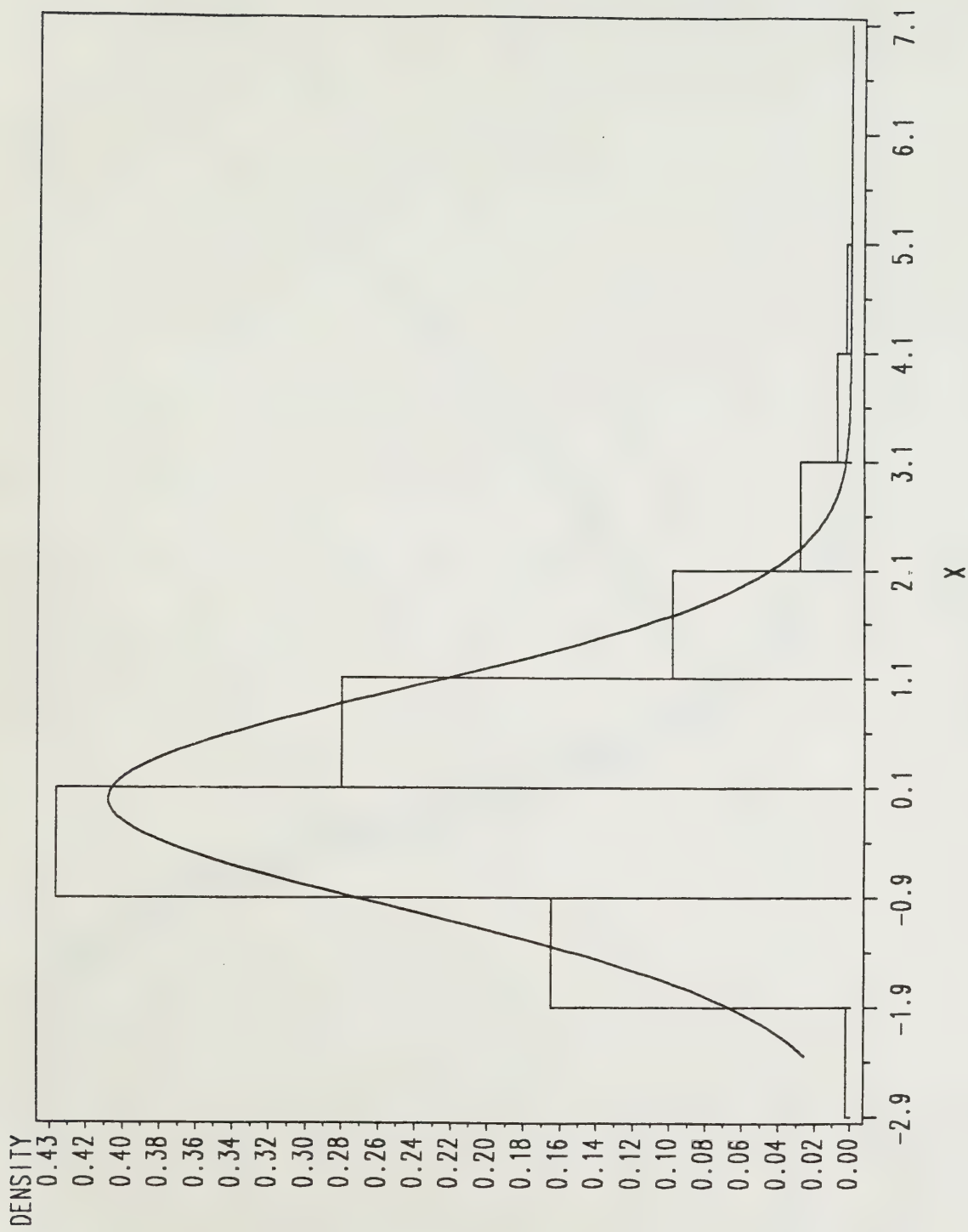
1. Histogram of standardized data and superimposed Normal(0,1) PDF, produced by %HIST Macro.
2. ECDF of standardized data and superimposed Normal(0,1) CDF, produced by %ECDF Macro.
3. Q-Q plot comparing unstandardized data to Normal(0,1) distribution, produced by %QQ Macro.
4. P-P plot comparing standardized data to Normal(0,1) distribution, produced by %PP Macro.

\* The data - observations of systolic blood pressure for 5802 Canadians - are from Statistics Canada's 1978/79 Canada Health Survey.



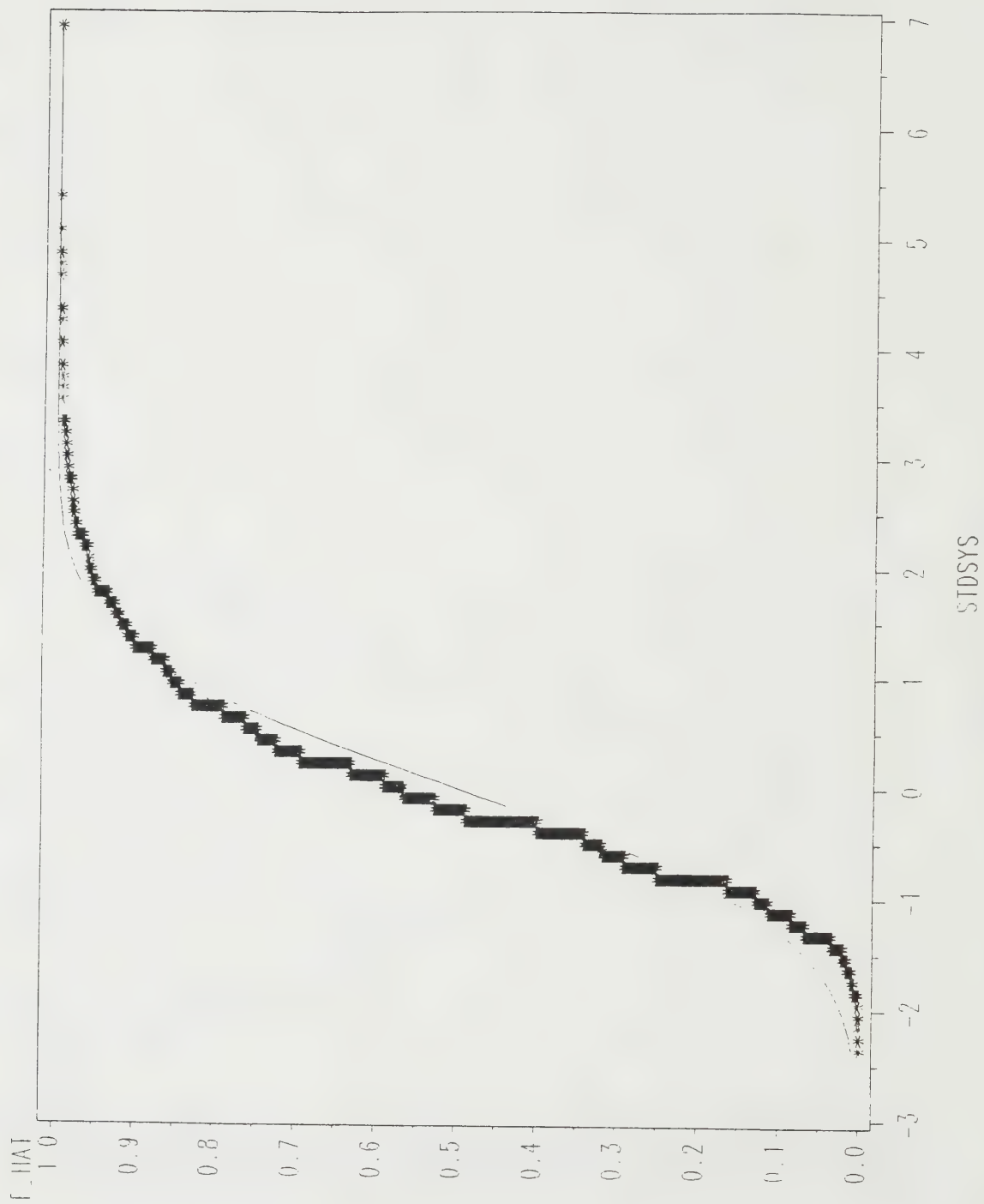
# SYSTOLIC BLOOD PRESSURE

MALES AND FEMALES



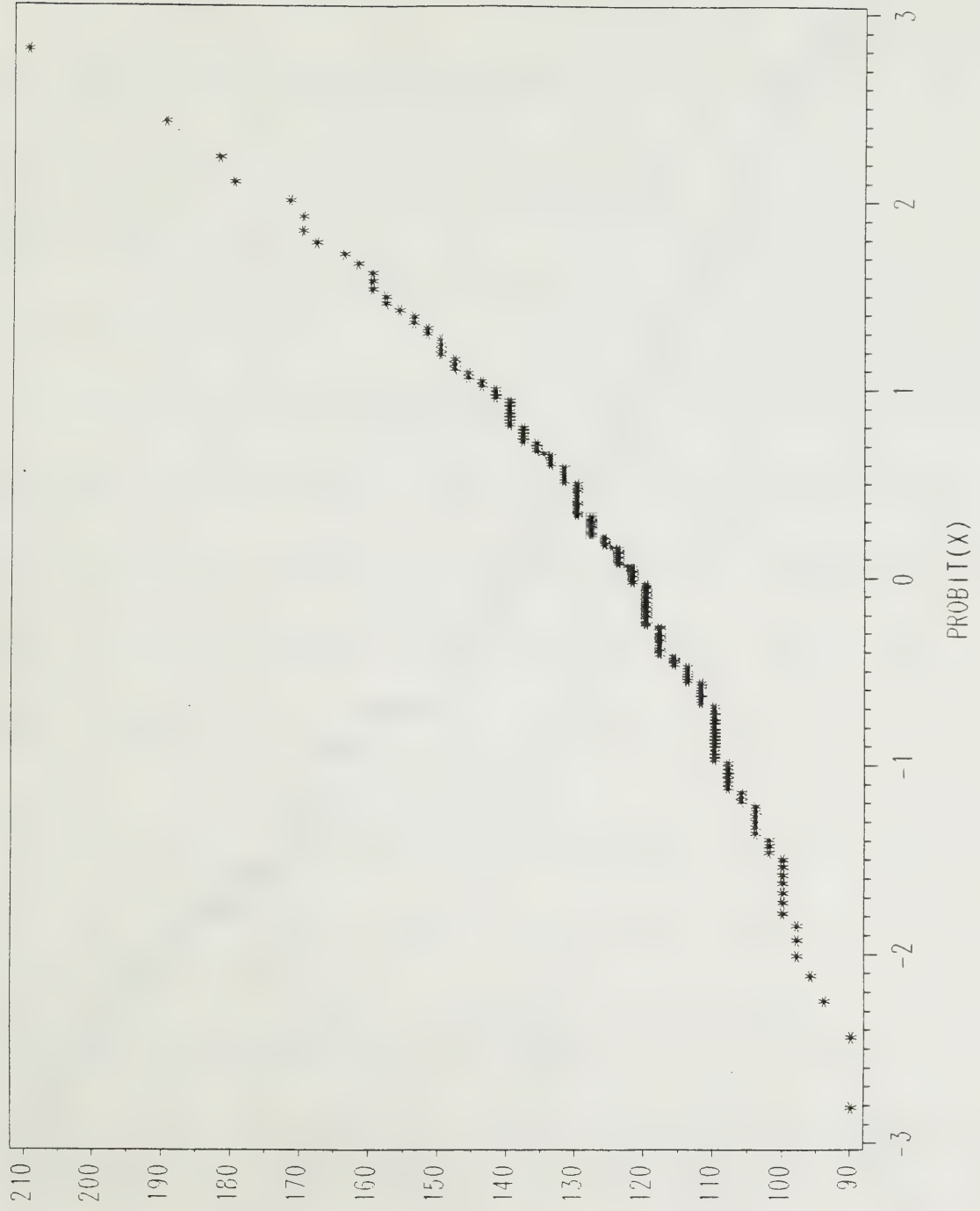
# SYSTOLIC BLOOD PRESSURE

## MALES AND FEMALES

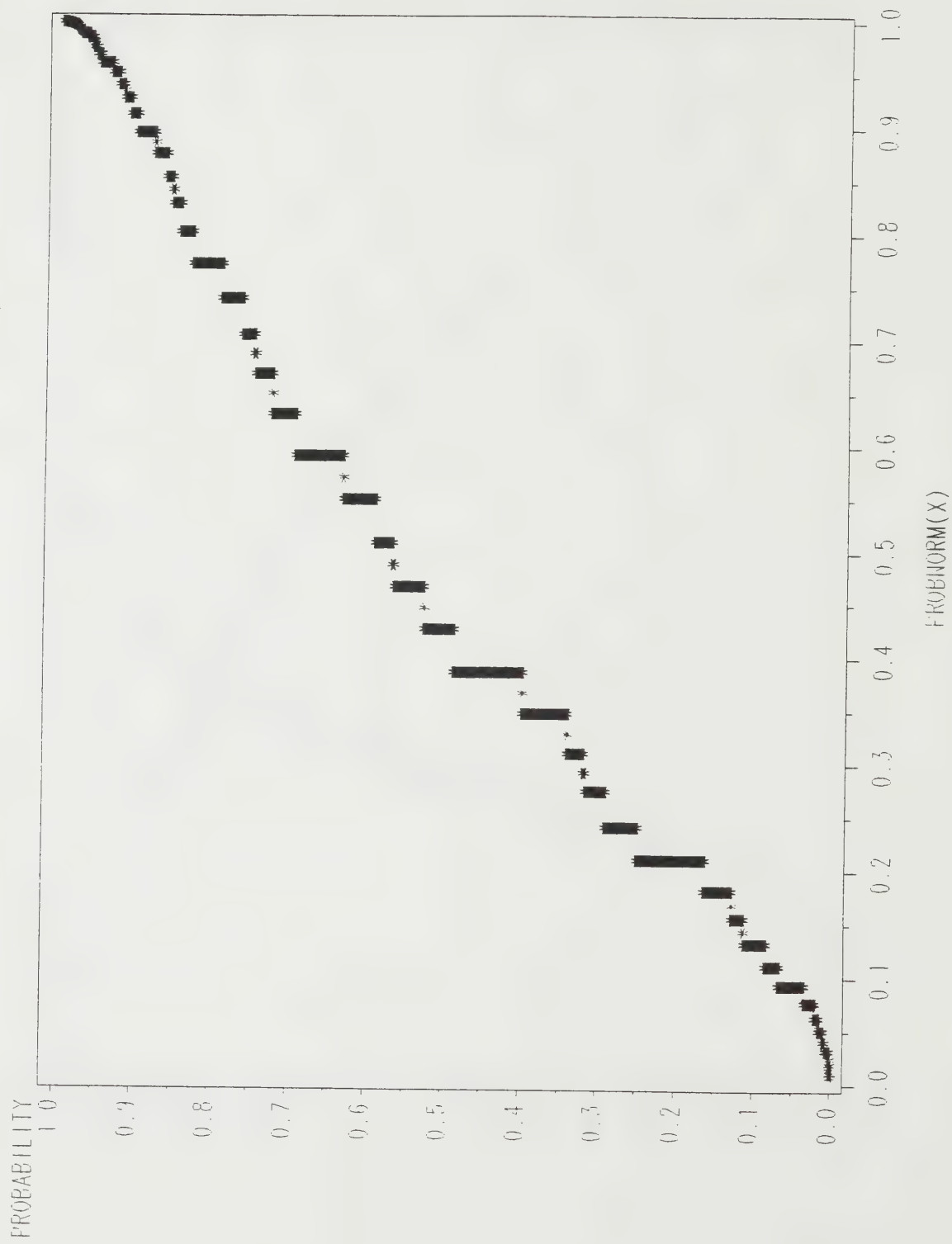


# SYSTOLIC BLOOD PRESSURE

MALES AND FEMALES



# SYSTOLIC BLOOD PRESSURE - MALES & FEMALES



ANALYTICAL STUDIES BRANCH --- RESEARCH PAPER SERIES

No.

1. BEHAVIOURAL RESPONSE IN THE CONTEXT OF SOCIO-ECONOMIC MICROANALYTIC SIMULATION by Lars Osberg
2. UNEMPLOYMENT AND TRAINING by Garnett Picot
3. HOMEMAKER PENSIONS AND LIFETIME REDISTRIBUTION by Michael C. Wolfson
4. MODELLING THE LIFETIME EMPLOYMENT PATTERNS OF CANADIANS by Garnett Picot
5. JOB LOSS AND LABOUR MARKET ADJUSTMENT IN THE CANADIAN ECONOMY by Garnett Picot and Ted Wannell
6. A SYSTEM OF HEALTH STATISTICS: Toward a New Conceptual Framework for Integrating Health Data by Michael C. Wolfson
7. A PROTOTYPE MICRO-MACRO LINK FOR THE CANADIAN HOUSEHOLD SECTOR by Hans J. Adler and Michael C. Wolfson
8. NOTES ON CORPORATE CONCENTRATION AND CANADA'S INCOME TAX by Michael C. Wolfson
9. THE EXPANDING MIDDLE: Some Canadian Evidence on the Deskillling Debate by John Myles
10. THE RISE OF THE CONGLOMERATE ECONOMY by Jorge Niosi
11. ENERGY ANALYSIS OF CANADIAN EXTERNAL TRADE: 1971 and 1976 by K.E. Hamilton
12. NET AND GROSS RATES OF LAND CONCENTRATION by Ray D. Bollman and Philip Ehrensaft
13. CAUSE-DELETED LIFE TABLES FOR CANADA (1921 TO 1981): An Approach Towards Analysing Epidemiologic Transition by Dhruva Nagnur and Michael Nagrodski
14. THE DISTRIBUTION OF THE FREQUENCY OF OCCURRENCE OF NUCLEOTIDE SUBSEQUENCES, BASED ON THEIR OVERLAP CAPABILITY by Jane F. Gentleman and Ronald C. Mullin
15. IMMIGRATION AND THE ETHNOLINGUISTIC CHARACTER OF CANADA AND QUEBEC by Réjean Lachapelle
16. INTEGRATION OF CANADIAN FARM AND OFF-FARM MARKETS AND THE OFF-FARM WORK OF WOMEN, MEN AND CHILDREN by Ray D. Bollman and Pamela Smith

17. WAGES AND JOBS IN THE 1980s: Changing Youth Wages and The Declining Middle by J. Myles, G. Picot and T. Wannell
18. A PROFILE OF FARMERS WITH COMPUTERS by Ray D. Bollman
19. MORTALITY RISK DISTRIBUTIONS: A Life Table Analysis by Geoff Rowe
20. INDUSTRIAL CLASSIFICATION IN THE CANADIAN CENSUS OF MANUFACTURES: Automated Verification Using Product Data by John S. Crysdale
21. CONSUMPTION, INCOME AND RETIREMENT by A.L. Robb and J. B. Burbridge
22. JOB TURNOVER IN CANADA'S MANUFACTURING SECTOR by John R. Baldwin and Paul K. Gorecki
- 23A. FIRM ENTRY AND EXIT IN THE CANADIAN MANUFACTURING SECTOR by John R. Baldwin and Paul K. Gorecki
24. MAINFRAME SAS ENHANCEMENTS IN SUPPORT OF EXPLORATORY DATA ANALYSIS by Richard Johnson and Jane F. Gentleman
25. DIMENSIONS OF LABOUR MARKET CHANGE IN CANADA: Intersectoral Shifts, Job and Worker Turnover by John R. Baldwin and Paul K. Gorecki

For further information, contact the Chairperson, Publication Review Committee, Analytical Studies Branch, R.H. Coats Bldg., 24th Floor, Statistics Canada, Tunney's Pasture, Ottawa, Ontario K1A 0T6.



